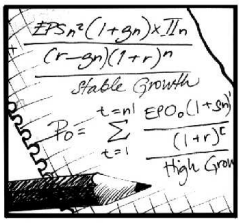

Presidentti-gallup

Pekka Alestalo
Matematiikan laitos, Teknillinen korkeakoulu

Gallup-uutinen

Presidentinvaalit lähestyvät. Vaalikampanjoinnin lisäksi kiihtyy myös tiedotusvälineiden gallup-huuma, kun tv, radio ja lehdistö kyllästävät meidät uutisilla, joiden yleinen kaava on seuraava:

"Ehdokas X johtaa kilpaa x % kannatuksella, toisena on ehdokas Y, jota kannattaa y % haastatelluista, ja ehdokas Z näyttää putoavan z % kannatuksellaan vaalin toiselta kierrokselta. Kyselyä varten haastateltiin n. 2 000 äänioikeutettua ja tulosten virhe on ±2 %-yksikköä."

Uutinen on niin tavallinen aikaisemmista vaaleista, ettei ehkä heti tule kiinnittäneeksi huomiota sen esittämän väitteen mahdottomuuteen. Koska äänioikeutettuja on Suomessa n. 4 miljoonaa, on kyselyyn osallistuneiden osuus kaikista mahdollisista äänestäjistä vain 0,05 %; ei kai näin pienen osuuden perusteella voi päätellä mitään todellisista kannatusluvuista. Vai voiko?

Pieni ajatusleikki lienee paikallaan. Pahimmassa tapauksessa ehdokkaalla X on koko maassa vain 2 000 kannattajaa, mutta juuri he sattuivat tulemaan valituiksi gallupiin. Haastattelijalta on esimerkiksi voinut mennä puhelinluettelo ja erään puolueen jäsenrekisteri sekaisin. Mutta vaikka haastateltujen valinta tehtäisiin kuinka huolellisesti ja "satunnaisesti" tahansa, on tällainen tulos kuitenkin periaatteessa mahdollinen. Tästä seuraa heti, ettei uutisen väite voi kirjaimellisesti ottaen pitää paikkansa.

Tarkkaavainen lukija on tietysti jo tässä vaiheessa huomannut, ettei yllä kuvattu tilanne ole kuitenkaan kovin *todennäköinen*. Itse asiassa kyseiselle todennäköisyydelle saadaan arvio

$$\frac{2\,000}{4\,000\,000}\cdot\frac{1\,999}{3\,999\,999}\cdots\frac{2}{3\,998\,002}\cdot\frac{1}{3\,998\,001}\approx 4{\cdot}10^{-7469}.$$

Käytännössä se on siis jotakuinkin mahdotonta. Ratkaisu uutisen sisältämään ristiriitaan piileekin siinä, että väitteestä on jätetty pois tämä todennäköisyyksiä koskeva osuus, jonka voisi arkikielellä ilmaista esimerkiksi muodossa "2 % virheraja on tosi 95 % todennäköisyydellä". Täsmällisemmässä kielessä sanotaan mieluummin, että väite on tosi 95 % luottamustasolla.

Seuraavassa on tarkoitus lyhyesti selvittää sitä, miten luvut 2 000, ±2 % ja 95 % liittyvät toisiinsa.

Otanta

Gallupissa haastateltavien henkilöiden valinta on esimerkki otannasta. Jatkossa ei kiinnitetä huomiota sii-

hen, millä tavalla nämä henkilöt pitäisi valita, jotta tulos olisi jossakin mielessä satunnainen. Oletamme esimerkiksi, että haastateltavat valitaan selaamalla puhelinluetteloita. (Mieti, mitä ongelmia tähän liittyy!)

Seuraavaksi yksinkertaistamme tilannetta olettamalla, että ehdokkaita on vain kaksi (tai että tutkitaan vain yhden ehdokkaan kannatusta suhteessa muihin). Voimme unohtaa vaaliteeman ja tarkastella otantaa, jossa sinisiä ja punaisia palloja sisältävästä laatikosta otetaan yksi pallo kerrallaan ja kirjataan sen väri. Tilanteen matemaattinen käsittely yksinkertaistuu, jos jokainen kirjattu pallo palautetaan laatikkoon, koska tällöin eriväristen pallojen suhteellinen osuus on jokaisessa toistossa sama. Jos palloja on yhteensä 4 000 000 ja niistä valitaan 2 000, ei tällä erolla ole mitään käytännön merkitystä. (Käytämme siis "otanta ilman takaisinpanoa"-menetelmän sijasta "otantaa takaisinpanolla")

Tarkastellaan tilannetta, jossa laatikossa on N palloa, joista P kpl punaisia ja S kpl sinisiä; tällöin $P+S=N$. Olkoon $p = P/N$ = laatikon punaisten pallojen suhteellinen osuus, $0 \le p \le 1$, ja vastaavasti $s = S/N = 1-p$ = sinisten pallojen suhteellinen osuus. Ongelmana on se, että kaikki luvut P, S, p, s ovat meille etukäteen tuntemattomia, mutta yritämme saada niistä tietoa valitsemalla laatikosta pallon n kertaa ja palauttamalla sen värin kirjaamisen jälkeen takaisin laatikkoon. Tällöin n toiston jälkeen punaisia palloja on todennäköisimmin saatu n. np ja sinisiä palloja n. ns kappaletta; täsmällisemmin sanottuna nämä luvut kuvaavat punaisten ja sinisten pallojen havaintokertojen odotusarvoja kyseisessä otannassa:

$$P(\text{1 pun. pallo}) \cdot 1 + P(\text{2 pun. palloa}) \cdot 2 + \cdots + P(n \text{ pun. palloa}) \cdot n = np.$$

Havaittujen punaisten pallojen lukumäärän varianssi voidaan laskea kaavalla

$$P(\text{1 pun. pallo}) \cdot (1-np)^2 + P(\text{2 pun. palloa}) \cdot (2-np)^2 + \cdots + P(n \text{ pun. palloa}) \cdot (n-np)^2 = nps.$$

Todennäköisyyksien P laskeminen ja tulosten tarkistaminen jääköön lukijalle, mutta asia löytyy myös joistakin lukiokirjoista binomijakauman kohdalta.

Selitys

Mutta miten näiden tietojen perusteella saadaan haettu yhteys? Tulemme nyt tämän kirjoitukseen vaikeimpaan kohtaan, joka tunnetaan todennäköisyyslaskennassa nimellä "Keskeinen raja-arvolause". Sen mukaan esimerkiksi otantaan liittyviä todennäköisyyksiä voidaan laskea sopivasti skaalatun normaalijakauman avulla: Kun n on kohtuullisen suuri, mutta selvästi pienempi kuin N, noudattaa esim. punaisten pallojen havaittu lukumäärä likimäärin sellaista normaalijakaumaa, jonka odotusarvo on $\mu = np$ ja keskihajonta $\sigma = \sqrt{nps}$. En käsittele raja-arvolausetta sen tarkemmin kuin mainitsemalla, että idealisoidussa tapauksessa $N = \infty$ rajatapaus $n \to \infty$ antaa täsmälleen oikean tuloksen kaikkiin tilannetta koskeviin todennäköisyyksiin.

Huomautettakoon, että tietokoneiden aikakaudella normaalijakauma-approksimaation käyttäminen ei ole välttämätöntä, sillä laskun lopputulos saadaan myös numeerisesti suoraan binomijakaumasta. Tällöin kaavan (1) antama riippuvuus jää kuitenkin epäselväksi.

Olkoon siis X = punaisten pallojen havaittu lukumäärä. Yllä mainittu skaalaaminen tarkoittaa sitä, että lauseke (= satunnaismuuttuja)

$$Y = \frac{X - np}{\sqrt{nps}}$$

on normaalijakautunut odotusarvolla $\mu = 0$ ja keskihajonnalla $\sigma = 1$; sen arvoja voidaan siis tutkia esim. MAOL-taulukon avulla. Haluaisimme päätellä, että otannan perusteella $p \approx X/n$ on havaittujen punaisten pallojen suhteellinen frekvenssi, mutta normaalijakaumaa käyttämällä voimme arvioida, kuinka luotettava näin saatu tulos on. Yleisesti käytetty "luotettavuuden" kriteeri on 95 % luottamustaso; vaaditun normaalijakauman osan pitäisi kattaa 95 % kaikista mahdollisuuksista. Jos siis haluamme väittää, että otannan tulos poikkeaa luvusta p korkeintaan a %-yksikköä 95 % luottamustasolla, niin vaatimus kuuluu:

$$P\{|X/n - p| \le a/100\} = 0{,}95.$$

Ehto, jonka todennäköisyyttä tutkimme, tulee muotoon $|X - np| \le na/100$, josta edelleen

$$|Y| = \frac{|X-np|}{\sqrt{nps}} \le \frac{na}{100\sqrt{nps}} = \sqrt{\frac{n}{ps}} \cdot \frac{a}{100}.$$

Koska Y noudattaa tavallista normaalijakaumaa, on ehdon $|Y| \le t$ toteutumisen todennäköisyys muotoa $\Phi(t) - \Phi(-t) = \Phi(t) - (1 - \Phi(t)) = 2\Phi(t) - 1$, missä Φ on vanha tuttu normaalijakauman kertymäfunktio.

Laskumme alkaa nyt olla loppuvaiheessa. Yhtälöstä

$$2\Phi\left(\sqrt{\frac{n}{ps}}\frac{a}{100}\right) - 1 = 0{,}95$$

saadaan ensin

$$\Phi\left(\sqrt{\frac{n}{ps}}\frac{a}{100}\right) = 0{,}975,$$

jolloin taulukon perusteella

$$\sqrt{\frac{n}{ps}}\frac{a}{100} \approx 1{,}96 \text{ eli } a/100 \approx \frac{2\sqrt{ps}}{\sqrt{n}}.$$

Luvut p, s ovat tuntemattomia, mutta koska $ps = p(1-p) \in [0, 1/4]$ aina, voimme varmuuden vuoksi käyttää maksimiarvoa $ps = 1/4$, jolloin otamme huomioon pahimman mahdollisen tilanteen. Näin saamme lopulta arvion

$$a \approx \frac{100}{\sqrt{n}}. \tag{1}$$

Ja sitten vain kokeillaan: Jos $n = 2\,000$, niin saadaan $a \approx 2{,}24$ %-yksikköä. Tämä arvio on siis laskettu 95 % luottamustasolla. Se on siinä!

Tehtävä: (i) Kuinka suuri n antaa tarkkuuden $a \approx 1$ %-yksikkö (luottamustasolla 95 %)?
(ii) Kuinka suuri n antaa tarkkuuden $a \approx 2$ %-yksikköä (luottamustasolla 99 %)?

Tommi Sottinen luki kirjoituksen läpi ja oikaisi muutaman väärinkäsityksen. Kiitokset!